

Changellenge CUP IT 2023

Секция Data Science

Задача “Естественный отбор”



Команда $O(N^2)$



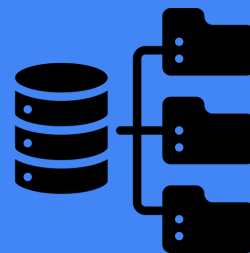
Ожерельев Виктор



Романенко Никита



Тенишев Никита



GitHub репозиторий

Задача: ранжирование комментариев с помощью ML

Знакомство с данными и
разведочный анализ
данных



Предобработка



Генерация новых
признаков



Выбор модели



Результаты. Выводы



Разведочный анализ данных

Пост в среднем
содержит

7 слов




Самый
популярный язык

En

Много постов о



Текст содержит

/xa0, , ,
, [https://...](#),
@,
...

В самом значимом
комментарии наибольшее
количество слов



Работа с текстом



Чистка текста и приведение к нормальному виду



Сходство частей речи в посте и комментарии



Количественная оценка текстов



Процент слов в комментарии относительно группы

Когда ждёшь вычисления
эмбедингов



Выбор модели

Существуют три дракона: XGBoost, LightGBM и CatBoost, но мы выбрали...



LightGBM

- + эффективно по времени
- дает хороший результат
- + занимает меньше памяти



CatBoost

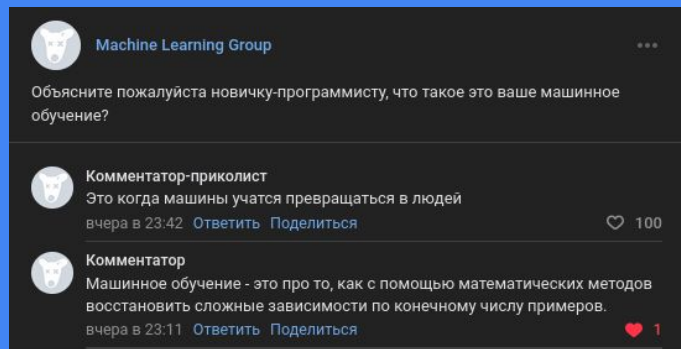
- + дает лучший результат
- дольше обучается
- занимает много памяти

Как наше решение поможет бизнесу

Комментарии ранжируются по дате их написания или количеству лайков

Семантика топовых комментариев не совпадает с семантикой поста

Внедряем наше решение



Добавляем рекламу в комментарии

Пользователи дольше залипают на постах, читая полезные комментарии

В топе оказываются комментарии, которые дополняют и/или раскрывают тему поста



Как наше решение поможет пользователю и бизнесу

Пользователь



Хороший отзыв и
оформление подписки

Пользователь получает то,
что хотел

Выскакивает подсказка с правильным
написанием релевантного текста и
пробной подпиской

Пользователь долго думает и
пишет нерелевантный текст

Пользователь пишет
комментарий, который должен
оказаться в топе

Подписка
Comment+

Идеи применения модели:

Детекция автоматически
сгенерированных
комментариев и/или
созданных со злым умыслом 1

Нахождение:
Запрещенных услуг
Мошенничества
Рекламы



Персональное ранжирование
комментариев для каждого
пользователя 2

Увеличение:
Времени просмотра ленты
Увлеченности пользователя



Идеи применения модели:

3

Question-answering:
ранжирование самых
релевантных комментариев
под постом, когда человек
пишет вопрос.



Увеличение:
популярность и
эффективность
комментариев



4

Спам-коммент:
определение спам-
комментария, который может
быть релевантным под постом,
но имеет мошенническую
основу.



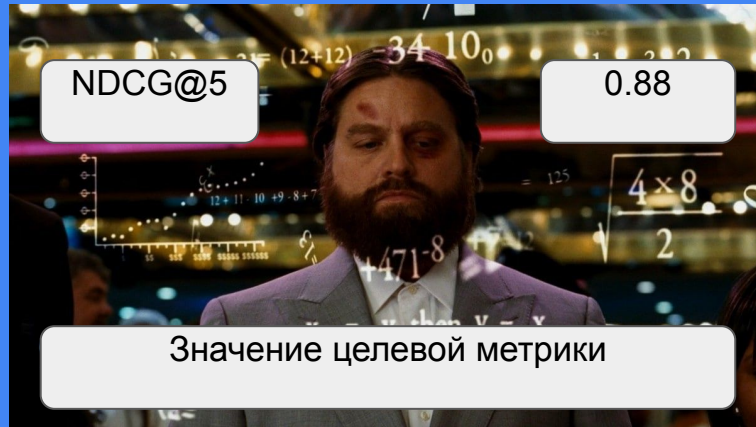
Уменьшение:
мошенничество
спам контент





Спасибо за внимание !

Приложение №1



Основной алгоритм ранжирования:
CatBoostRanker

Использованы стат. тесты:

- HSD критерий Тьюки
- хи-квадрат

Приложение №2



Доступные вычислительные
мощности:

- GeForce RTX 3050ti laptop
- intel core i5 11400H
- Nvidia Tesla T4 (Google Colab)

Bert:

- для определения токсичности
- для построения эмбедингов



Приложение №3

Идеи

```
graph TD; A[Идеи] --> B[Оценка темы поста и комментария]; A --> C[Количество различных слов относительно других комментариев];
```

Оценка темы поста и комментария

Количество различных слов
относительно других комментариев